УДК 519.7

БЫСТРЫЙ ВЫБОР КОЭФФИЦИЕНТОВ РАЗМЫТОСТИ ЯДЕРНОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ НЕЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

© **А. В. Лапко**[1, 2], **В. А. Лапко**[1, 2]

[1] *Институт вычислительного моделирования СО РАН,*
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Сибирский государственный университет науки и технологий*
им. академика М. Ф. Решетнева,
660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Предлагается методика быстрого выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности двухмерной случайной величины с независимыми компонентами. Исследуемые законы распределения принадлежат семейству одномодальных и симметричных плотностей вероятностей. Обосновывается возможность их оптимизации на основе анализа асимптотических выражений средних квадратических отклонений компонент двухмерной случайной величины. Каждая компонента характеризуется оптимальным коэффициентом размытости ядерных функций, который зависит от нелинейного функционала плотности вероятности. Установлена его функциональная зависимость от коэффициента контрэксцесса одномерной случайной величины. Эффективность предлагаемой методики подтверждается результатами аналитических исследований.

Ключевые слова: ядерная оценка плотности вероятности, быстрый алгоритм выбора коэффициента размытости, независимые случайные величины, коэффициент контрэксцесса, одномодальные симметричные законы распределения.

DOI: 10.15372/AUT20220104

Введение. При синтезе алгоритмов обработки статистических данных в условиях априорной неопределённости вида исследуемых закономерностей широко используются непараметрические оценки плотности вероятности $\bar{p}(x)$ типа Розенблатта — Парзена. Традиционные методы оптимизации $\bar{p}(x)$ основаны на выборе коэффициентов размытости ядерных функций из условия минимума оценки среднего квадратического отклонения $\bar{p}(x)$ от плотности вероятности $p(x)$ либо максимума функции правдоподобия [1–10]. Реализация указанных методов сопряжена с большими временны́ми затратами, которые возрастают с увеличением объёма статистических данных. Поэтому в настоящее время значительное внимание уделяется решению проблемы быстрого выбора коэффициентов размытости ядерных оценок плотности вероятности $\bar{p}(x)$ [11–22]. Предлагаемый подход использует результаты анализа формулы оптимального коэффициента размытости, минимизирующего асимптотическое выражение среднего квадратического отклонения $\bar{p}(x)$ от $p(x)$. Основная составляющая формулы оптимального коэффициента размытости — нелинейный функционал от второй производной $p(x)$. Установлено, что исследуемый функционал является константой для ряда семейств одномодальных и симметричных плотностей вероятностей. Получены зависимости его значений от количественных характеристик законов распределения случайных величин [18, 19, 22, 23].

В данной работе методика быстрого выбора коэффициентов размытости ядерных функций развивается на задачу оптимизации непараметрической оценки плотности вероятности двухмерной случайной величины с независимыми компонентами.

Непараметрическая оценка плотности вероятности независимых случайных величин и её свойства. Имеется выборка $V = (x^i,\ i = \overline{1,n}\,)$ двухмерной случайной величины $x = (x_1, x_2)$ с априори неизвестной плотностью вероятности $p(x) = p(x_1)p(x_2)$.

Для оценивания плотности вероятности $p(x)$ воспользуемся непараметрической статистикой

$$\bar{p}(x) = \bar{p}(x_1)\,\bar{p}(x_2),$$

где

$$\bar{p}(x_v) = \frac{1}{nc_v}\sum_{i=1}^{n}\Phi\Big(\frac{x_v - x_v^i}{c_v}\Big). \tag{1}$$

Ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\Phi(u_v) = \Phi(-u_v); \qquad 0 \leqslant \Phi(u_v) < \infty; \qquad \int \Phi(u_v)\,du_v = 1; \qquad \int u^2\Phi(u_v)\,du_v = 1;$$

$$\int u^m\Phi(u_v)\,du_v < \infty, \qquad 0 \leqslant m < \infty, \quad v = 1,2.$$

Здесь и далее бесконечные пределы интегрирования опускаются.

В работе [24] при $\Phi(u) = \Phi(u_v)$, $v = 1,2$, получено асимптотическое выражение среднего квадратического отклонения $\bar{p}(x_1)\,\bar{p}(x_2)$ от $p(x_1)\,p(x_2)$:

$$W_{12}(c_1,c_2) = \frac{\|\Phi(u)\|^2\|p(x_2)\|^2}{nc_1} + \frac{c_1^4}{4}\,\|p(x_2)\|^2\|p^{(2)}(x_1)\|^2 +$$

$$+ \frac{\|\Phi(u)\|^2\|p(x_1)\|^2}{nc_2} + \frac{c_2^4}{4}\,\|p(x_1)\|^2\|p^{(2)}(x_2)\|^2 +$$

$$+ \frac{(\|\Phi(u)\|^2)^2}{n^2c_1c_2} + \frac{c_1^2c_2^2}{2}\int p(x_1)p^{(2)}(x_1)\,dx_1\int p(x_2)p^{(2)}(x_2)\,dx_2, \tag{2}$$

где $p^{(2)}(x_v)$ — вторая производная $p(x_v)$ по x_v; $\|\Phi(u)\|^2 = \int\Phi^2(u)\,du$; $\|p^{(2)}(x_v)\|^2 = \int(p^{(2)}(x_v))^2\,dx_v$; $\|p(x_v)\|^2 = \int p^2(x_v)\,dx_v$.

Определение оптимальных коэффициентов размытости c_1, c_2 из условия минимума выражения (2) связано с решением двух уравнений пятой степени от искомых параметров, которые не имеют аналитического решения. Поэтому в качестве коэффициентов размытости c_1, c_2 в статистике $\bar{p}(x_1)\,\bar{p}(x_2)$ будем использовать значения $\bar{c}_1^*$, $\bar{c}_2^*$, минимизирующие каждую из двух первых пар слагаемых выражения (2) соответственно. Возможность предлагаемого подхода подтверждается результатами вычислительных экспериментов. Например, для нормальных законов распределения $N(0;\sigma_1)$ и $N(0;\sigma_2)$ отношение $W_{12}(\bar{c}_1^*,\bar{c}_2^*)/W_{12}(c_1^*,c_2^*) \in [1{,}0056;1{,}0075]$ при $\sigma_1 = 1$, $\sigma_2 \in [0{,}5;3]$ и $n \in [100;500]$.

Из условия минимума асимптотического выражения среднего квадратического отклонения $\bar{p}(x_v)$ от $p(x_v)$:

$$W(\bar{c}_v) \sim (n\,\bar{c}_v)^{-1}\,\|\Phi(u)\|^2 + \frac{\bar{c}_v^4\,\|p^{(2)}(x_v)\|^2}{4}, \tag{3}$$

оптимальное значение коэффициента размытости c_v ядерных функций статистики (1) определяется формулой

$$\bar{c}_v^* = \Big(\frac{\|\Phi(u)\|^2}{n\,\|p^{(2)}(x_v)\|^2}\Big)^{1/5}, \qquad v = 1, 2. \tag{4}$$

Обозначим первые четыре слагаемые выражения (2) через

$$W'_{12}(c_1, c_2) = W(c_1) + W(c_2),$$

где $W(c_1)$, $W(c_2)$ соответствуют средним квадратическим отклонениям $\bar{p}(x_1)$, $\bar{p}(x_2)$.

При оптимальных значениях $\bar{c}_v^*$, $v = 1, 2$, основная составляющая критерия (2) принимает вид

$$W'_{12}(\bar{c}_1^*, \bar{c}_2^*) = \frac{5}{4}\sum_{v=1}^{2}\Big(\Big(\frac{\|\Phi(u)\|^2}{n}\Big)^4 \|p^{(2)}(x_v)\|^2\Big)^{1/5}. \tag{5}$$

Нетрудно заметить, что формула (4) после несложных преобразований запишется как

$$\bar{c}_v^* = \beta_v \sigma_v n^{-1/5}. \tag{6}$$

В формуле (6) значение

$$\beta_v = \Big(\frac{\|\Phi(u)\|^2}{\lambda_v}\Big)^{1/5}, \tag{7}$$

где

$$\lambda_v = \sigma_v^5\,\|p^{(2)}(x_v)\|^2, \qquad v = 1, 2, \tag{8}$$

является константой для семейства одномодальных плотностей вероятностей. Этим семействам соответствуют, например, гауссовский, Лапласа, логистический и экспоненциальный законы распределения [18, 19, 22, 23].

Возникает задача оценивания зависимости константы λ_v от количественных характеристик восстанавливаемой плотности вероятности. Результаты её решения позволят повысить вычислительную эффективность методики быстрого выбора коэффициентов размытости ядерных функций при построении непараметрической оценки плотности вероятности независимых случайных величин.

С учётом вышеизложенного обозначим через $\bar{\lambda}_v$ оценку константы λ_v, а соответствующий ей коэффициент размытости ядерных функций (4) статистики (1) как $\bar{\bar{c}}_v^*$. Определим отношение средних квадратических отклонений $\bar{p}(x_v)$ от $p(x_v)$ в принятых условиях в виде

$$\bar{R} = W'_{12}(\bar{\bar{c}}_1^*, \bar{\bar{c}}_2^*)/W'_{12}(\bar{c}_1^*, \bar{c}_2^*).$$

Анализ отношения $\bar{R}$ для одномодальных и симметричных плотностей вероятностей двухмерной случайной величины. Пусть законы распределений случайных величин являются одномодальными и симметричными, например гауссовские, логистические и законы распределения Стьюдента. Для этого семейства плотностей вероятностей в работе [22] определена функциональная зависимость между константой λ_v и коэффициентом контрэксцесса δ_v, которая имеет вид

$$\bar{\lambda}_v = -2{,}185\delta_v + 1{,}4635 \tag{9}$$

при средней относительной ошибке аппроксимации $\bar{\rho}_v = 0{,}0365$, $v = 1, 2$.

Исследуем влияние $\bar{\rho}_v$, $v = 1, 2$, на аппроксимационные свойства статистики $\bar{p}(x_1)\,\bar{p}(x_2)$. Для этого проведём анализ отношения $\bar{R}$. Обозначим через $\alpha\lambda_v = (1 \pm \bar{\rho}_v)\lambda_v$ результат оценивания λ_v (8) с использованием её модели (9). В условиях принятой погрешности оценивания константы λ_v значение оптимального коэффициента размытости статистики (1) с учётом формул (7), (8) запишется как

$$\bar{\bar{c}}_v^{\,*} = \alpha_v^{-1/5}\,\bar{c}_v^{\,*}.$$

Определим среднее квадратическое отклонение $\bar{p}(x_v)$ от $p(x_v)$ при значении $\bar{\bar{c}}_v^{\,*}$:

$$W'_{12}(\bar{\bar{c}}_v^{\,*}) = K_{12}(v)\Big(\Big(\frac{\|\Phi(u)\|^2}{n}\Big)^4 \|p^{(2)}(x_v)\|^2\Big)^{1/5}, \qquad v = 1, 2, \tag{10}$$

где $K_{12}(v) = (4\alpha_v + 1)/(4\alpha_v^{4/5})$.

Тогда основная составляющая критерия (2) при значениях $\bar{\bar{c}}_v^{\,*}$, $v = 1, 2$, принимает вид

$$W'_{12}(\bar{\bar{c}}_1^{\,*}, \bar{\bar{c}}_2^{\,*}) = \sum_{v=1}^{2} K_{12}(v)\Big(\Big(\frac{\|\Phi(u)\|^2}{n}\Big)^4 \|p^{(2)}(x_v)\|^2\Big)^{1/5}. \tag{11}$$

Если ошибки оценивания λ_v равны нулю, то значения $\alpha_v = 1$, $v = 1, 2$, и критерии (5), (11) совпадают. Отмеченный вывод подтверждает корректность выполненных преобразований.

Предположим, что виды $p(x_1)$, $p(x_2)$ не различаются и принадлежат рассматриваемому семейству одномодальных и симметричных плотностей вероятностей. Тогда $\alpha_v = \alpha$, $v = 1, 2$, и коэффициент $K_{12}(v)$ в критерии (11) представляется в виде

$$K_{12} = \frac{4\alpha + 1}{4\alpha^{4/5}},$$

зависимость которого от значений α представлена на рис. 1.

Для этих условий вычислим отношение критериев (11) и (5):

$$\bar{R} = \frac{4\alpha + 1}{5\alpha^{4/5}}, \tag{12}$$

которое при ошибках оценивания $\bar{\rho}_v = 0$, $v = 1, 2$, равно единице. Зависимость $\bar{R}$ от параметра α представлена на рис. 2.

Если виды плотностей вероятностей $p(x_1)$, $p(x_2)$ из рассматриваемого семейства различаются, то анализ отношения $\bar{R}$ необходимо осуществлять с учётом общего вида критериев (5) и (11).

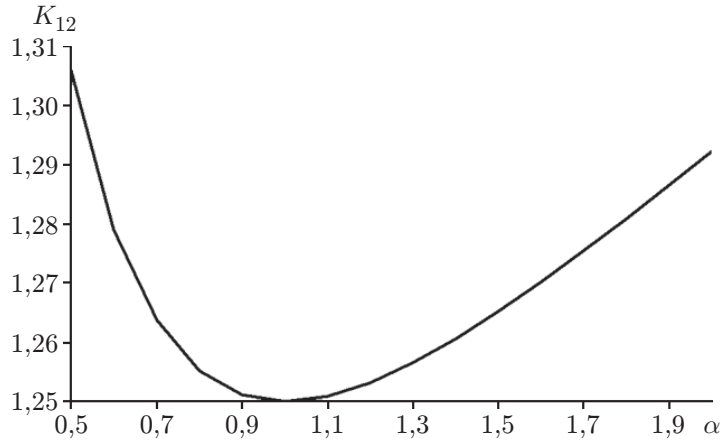

Рис. 1. Зависимость коэффициента K_{12} в критерии (11) от значений параметра α

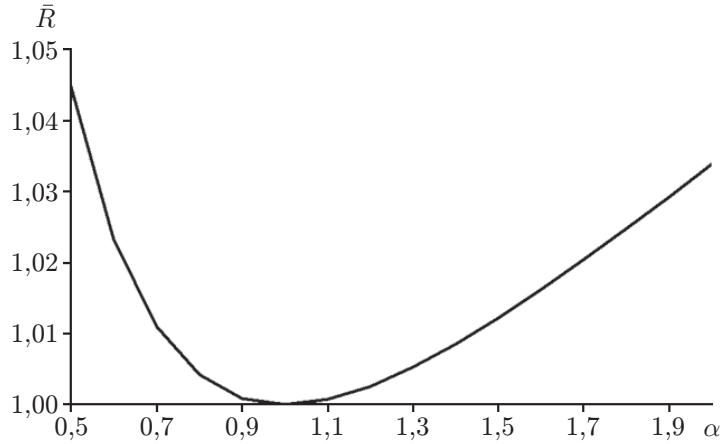

Рис. 2. Зависимость отношения $\bar{R}$ от значений параметра α

Исследуем зависимость отношения $\bar{R}$ от параметра α (см. рис. 2). Отношение $\bar{R}$ более чувствительно к уменьшению значений параметра α по сравнению с его увеличением. Отсюда следует, что ошибки, связанные с увеличением значений коэффициентов размытости c_1, c_2 ядерных функций в статистике $\bar{p}(x_1)\,\bar{p}(x_2)$, оказывают меньшее влияние на её аппроксимационные свойства по сравнению с уменьшением значений c_1, c_2. Например, для нормальных законов распределения $p(x_1)$ и $p(x_2)$ значения констант $\lambda_v = 0{,}212$, $v = 1, 2$, не зависят от параметров рассматриваемых плотностей вероятностей. В соответствии с моделью (9) и $\delta_v = 0{,}577$ значение $\bar{\lambda}_v = 0{,}202$ при относительной ошибке аппроксимации $\bar{\rho}' = |\lambda - \bar{\lambda}|/\lambda$ равно 0,047. В этих условиях $\alpha \in [0{,}953; 1{,}047]$, которым соответствуют значения $\bar{R} \in [1{,}000\,187; 1{,}000\,167]$.

Для логистических законов распределения $p(x_1)$, $p(x_2)$ значения констант λ_v равны 0,467, $v = 1, 2$. Их оценки при использовании модели (9) при $\delta_v = 0{,}488$ соответствуют значениям $\bar{\lambda}_v = 0{,}397$, $v = 1, 2$. В этих условиях относительная ошибка аппроксимации $\bar{\rho}' = 0{,}149$, параметр $\alpha \in [0{,}851; 1{,}149]$, а соответствующее им значение $\bar{R} \in [1{,}002\,152; 1{,}001\,502]$.

Исследуем отношение $\bar{R}$ для условий, когда плотности вероятностей $p(x_1)$ и $p(x_2)$ принадлежат рассматриваемому семейству, но их вид различается. Пусть $p(x_1)$, $p(x_2)$ соответствуют нормальной и логистической плотностям вероятности. При вычислении критериев (5) и (11) используется ядерная функция В. А. Епанечникова, для которой

$\|\Phi(u)\|^2 = 3/(5\sqrt{5}\,)$ [25]. Тогда с учётом вышеприведённых значений параметров $\bar{\rho}_v$, α_v, $v = 1, 2$, отношение $\bar{R} = 1{,}001$. Это отношение является константой и не зависит от значений $\bar{\rho}_v$, α_v, $v = 1, 2$, и параметров рассматриваемых законов распределения. Отметим, что значение $\bar{R}$ не зависит от объёма n статистических данных. Этот факт является ожидаемым, так как значения коэффициентов контрэксцесса случайных величин на данном этапе исследований считаются известными.

Заключение. Быстрые алгоритмы выбора коэффициентов размытости ядерных функций непараметрической оценки плотности вероятности типа Розенблатта — Парзена позволяют на порядки сократить временны́е затраты на её оптимизацию. Асимптотические свойства среднего квадратического отклонения двухмерной случайной величины с независимыми компонентами определяются в основном соответствующими им критериями. Поэтому выбор коэффициентов размытости исследуемой оценки плотности вероятности можно осуществлять из условий минимума средних квадратических отклонений её компонент. Получаемые оптимальные коэффициенты размытости ядерных функций зависят от нелинейных функционалов плотностей вероятностей компонент двухмерной случайной величины, для оценивания которых используются их коэффициенты контрэксцесса. Отношение средних квадратических отклонений непараметрических оценок плотностей вероятностей в условиях оценок коэффициентов размытости ядерных функций и их оптимальных значений не превышает величину 1,001. Это отношение сохраняется при различных вариантах сочетаний компонент двухмерных плотностей вероятностей независимых случайных величин и не определяется их параметрами.

Дальнейшее развитие предлагаемой методики состоит в исследовании влияния ошибок оценивания коэффициентов контрэксцесса на аппроксимационные свойства непараметрической оценки плотности вероятности независимых случайных величин.

СПИСОК ЛИТЕРАТУРЫ

1. **Rudemo M.** Empirical choice of histogram and kernel density estimators // Scand. Journ. Statist. 1982. N 9. P. 65–78.
2. **Bowman A. W.** A comparative study of some kernel-based non-parametric density estimators // Journ. Statist. Comput. Simul. 1982. **21**, Iss. 3–4. P. 313–327.
3. **Hall P.** Large-sample optimality of least squares cross-validation in density estimation // Ann. Statist. 1983. **11**, N 4. P. 1156–1174.
4. **Jiang M., Provost S. B.** A hybrid bandwidth selection methodology for kernel density estimation // Journ. Statist. Comput. and Simul. 2014. **84**, Iss. 3. P. 614–627. DOI: 10.1080/00949655.2012.721366.
5. **Dutta S.** Cross-validation revisited // Commun. Statist. Simul. and Comput. 2016. **45**, Iss. 2. P. 472–490. DOI: 10.1080/03610918.2013.862275.
6. **Heidenreich N.-B., Schindler A., Sperlich S.** Bandwidth selection for kernel density estimation: A review of fully automatic selectors // AStA Adv. Statist. Anal. 2013. **97**, N 4. P. 403–433. DOI: 10.1007/s10182-013-0216-y.
7. **Li Q., Racine J. S.** Nonparametric Econometrics: Theory and Practice. Princeton: Princeton University Press, 2007. 768 p.
8. **Duin R. P. W.** On the choice of smoothing parameters for Parzen estimators of probability density functions // IEEE Trans. Comput. 1976. **25**, Iss. 11. P. 1175–1179.
9. **Botev Z. I., Kroese D. P.** Non-asymptotic bandwidth selection for density estimation of discrete data // Methodol. and Comput. Appl. Probability. 2008. **10**, N 3. P. 435–451.

10. **Лапко А. В., Лапко В. А.** Анализ методов оптимизации непараметрической оценки плотности вероятности по коэффициенту размытости ядерных функций // Измерительная техника. 2017. № 6. С. 3–8.
11. **Варжапетян А. Г., Михайлова Е. Ю.** Методы выбора определяющих характеристик непараметрических алгоритмов идентификации моделей надёжности сложных систем по эксплуатационным данным // Вопросы кибернетики. 1982. Вып. 94. С. 77–87.
12. **Silverman B. W.** Density Estimation for Statistics and Data Analysis. London: Chapman & Hall, 1986. 175 p.
13. **Sheather S., Jones M.** A reliable data-based bandwidth selection method for kernel density estimation // Journ. Royal Statist. Soc. Ser. B. 1991. **53**, N 3. P. 683–690. DOI: 10.1111/j.2517-6161.1991.tb01857.x.
14. **Sheather S. J.** Density estimation // Statist. Sci. 2004. **19**, N 4. P. 588–597. DOI: 10.1214/088342304000000297.
15. **Terrell G. R., Scott D. W.** Oversmoothed nonparametric density estimates // Journ. Amer. Statist. Assoc. 1985. **80**, N 389. P. 209–214.
16. **Jones M. C., Marron J. S., Sheather S. J.** A brief survey of bandwidth selection for density estimation // Journ. Amer. Statist. Assoc. 1996. **91**. P. 401–407.
17. **Scott D. W.** Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons, 2015. 384 p.
18. **Лапко А. В., Лапко В. А.** Быстрый выбор коэффициентов размытости в многомерном непараметрическом алгоритме распознавания образов // Измерительная техника. 2019. № 8. С. 8–13. DOI: 10.32446/0368-1025it.2019-8-8-13.
19. **Лапко А. В., Лапко В. А.** Оценивание интеграла от квадрата производных симметричных плотностей вероятностей одномерных случайных величин // Метрология. 2020. № 1. С. 15–27. DOI: 10.32446/0132-4713.2020-1-15-27.
20. **Лапко А. В., Лапко В. А.** Оценивание нелинейного функционала от плотности вероятности при оптимизации непараметрических решающих функций // Измерительная техника. 2021. № 1. С. 14–20. DOI: 10.32446/0368-1025it.2021-1-14-20.
21. **Лапко А. В., Лапко В. А.** Анализ отношения средних квадратических отклонений ядерной оценки плотности вероятности в условиях независимых и зависимых случайных величин // Измерительная техника. 2021. № 3. С. 9–14. DOI: 10.32446/0368-1025it.2021-3-9-14.
22. **Лапко А. В., Лапко В. А.** Модифицированный алгоритм быстрого определения коэффициента размытости ядерной оценки плотности вероятности // Автометрия. 2020. **56**, № 6. С. 11–18. DOI: 10.15372/AUT20200602.
23. **Лапко А. В., Лапко В. А.** Быстрый алгоритм выбора коэффициентов размытости ядерных функций в непараметрической оценке плотности вероятности // Измерительная техника. 2018. № 6. С. 16–20. DOI: 10.32446/0368-1025it-2018-6-16-20.
24. **Лапко А. В., Лапко В. А.** Непараметрическая оценка плотности вероятности независимых случайных величин // Информатика и системы управления. 2011. **29**, № 3. С. 118–124.
25. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. **14**, № 1. С. 156–161.

Поступила в редакцию 20.08.2021
После доработки 06.12.2021
Принята к публикации 28.12.21